ChinaXiv合作期刊

多源数据融合视角下的大学生“消费–学业–社交”画像构建研究

黄泰华 [1]，张　涛 [1*]，王　磊 [2]
（1. 黑龙江大学 信息管理学院，哈尔滨　150080；2. 黑龙江大学 数据科学与技术学院，哈尔滨　150080）

摘　要：[目的/意义]挖掘高校学生数据构建学生画像，使高校管理过程中的学生形象具体化，利用数据分析手段深入了解学生需求，着力提升高校信息管理水平，推进管理和服务智能化。[方法/过程]基于高校管理和服务过程产生的多源数据，聚焦消费、学业和社交 3 类指标，利用 MySQL 和 SPSS 手段构建学生个体画像，利用 Python 中 sklearn 工具实现 K-means 聚类算法，构建学生群体画像，开展学生画像实证研究，并剖析学生画像的应用表征。[结果/结论]多源数据融合视角下的学生画像可以从个体和群体两个维度构建，个体画像表现直观，群体画像区分显著。可实现异常识别与预警、群体关注与引导和资源规划与调节等方面应用，有利于增加高校管理精度，提升学生获得感，为高校贫困助学、学业帮扶和心理干预等工作提供参考。

关键词：学生画像；消费分析；社交分析；学业分析；K-means 聚类；信息行为

中图分类号：G259.7　　　文献标识码：A　　　文章编号：1002-1248（2022）07-0076-12

引用本文：黄泰华，张涛，王磊. 多源数据融合视角下的大学生“消费–学业–社交”画像构建研究[J]. 农业图书情报学报，2022，34(7)：76-87.

chinaXiv:202303.10408v1

1 引　言

随着中国信息技术的发展，大数据技术正逐步应用于社会的各行各业，改善人们的学习、工作和生活，在此背景下，中国高校面临的内部结构和外部环境正在发生前所未有的深刻变化，学生管理工作中存在许多新情况、新问题、新挑战[1]。高校如何借助技术优势实现教育管理的多层面影响，已成为新时期高校教育管理研究实践的重要课题[2]。高校智慧校园的内生需求之一即是基于大数据分析实现校情研判并指引决策制定，与需求相悖的是高校内部的学生数据通常是海量的、异构的、复杂的，甚至是不完善的，多源数据融合为实现校园信息化治理提供了新的研究思路。此外，

收稿日期：2022–03–17

基金项目：黑龙江省教育科学“十四五”规划 2022 年度重点课题“后疫情时代高校线上教学效果影响因素实证分析”（GJB1422044）；2021 年度黑龙江省高等教育教学改革研究一般研究项目“跨学科人才培养背景下大学生数据素养教育评价体系研究”（SJGY20210251）

作者简介：黄泰华（1998– ），男，硕士，研究方向为用户画像。王磊（1976– ），男，研究员，研究方向为教育信息化建设与规划

*通信作者：张涛（1981– ），男，副教授，硕士生导师，研究方向为文本计算与数据分析。Email：zhangtao@hlju.edu.cn

chinaXiv:202303.10408v1

用户画像作为一种信息化的用户描述工具，在用户描述与建模上具有优势[3]。因此，将传统的高校管理经验与新时代的信息化手段相结合，建构高效能、信息化的教育管理体系，已成为新时期提升教育教学能效的关键基础，也为高校教育教学改革指明了方向。

2 相关研究

2.1 用户画像技术的相关研究

用户画像的概念最早由 A. Cooper 提出，意为“真实用户的虚拟代表”，侧重于探索用户的动机，是基于一系列真实数据的目标用户模型。为了更好地对学生数据进行深度挖掘，可以应用用户画像的研究方法，构建面向大学生的学生画像。在国外研究中，有部分学者将用户画像应用于图书馆管理工作中，识别图书馆用户的独特性质，进一步开发和改进当前服务并创建新服务以满足用户的需求[4]。有学者构建了基于数字画像的综合素质评价模型[5]。有学者提出了可视化的学习分析技术，构建了研究性学习学生画像[6]。有学者通过提出“精英模型”，对现有的学生画像完善拓展[7]。在上述研究中，数据挖掘的角度和手段在不断地创新。既有面向教学方面，实现学业预警；也有应用于消费方面，通过分析消费行为识别特征群体，实现贫困资助工作的有效开展；也有应用于心理评估方面，实现重点学生识别和关怀。

用户画像的构建方法主要包括基于用户行为、基于用户兴趣偏好、基于主题、基于人格特征与情绪 4 种方法，其应用领域大致涉及电子商务、健康医疗、旅游业、图书馆等领域。在教育领域的用户画像研究中，主要集中在基础教育研究，中国有关高等教育的用户画像研究尚处于起步阶段。根据现有文献来看，用户画像在高校管理中的应用研究主要包括教育教学管理、学生工作管理和图书馆管理 3 个方面。通过对国内外有关高校学生画像研究的内容梳理发现：①用户画像是一个新兴的研究领域，具备坚实的理论基础、成熟的研究方法和广泛的应用场景，但国内有关教育领域的相关研究较少，存在一定的研究空白；②在高校管理中用户画像研究中，多集中于图情管理领域，针对学生画像的研究多停留于数据分析层面，深层次的学生画像的构建及应用研究较少。

2.2 大学生行为分析的相关研究

大学生基础素质和知识水平较高，思想活跃，因此，从学生行为视角入手，在智慧育人的理念下，将高校学生的数据信息作为研究对象，探索大学生精准服务的新模式[8]，往往是专家学者开展高校教育教学体制研究的起点。国外研究中也常常利用学生行为数据以分析个人和学校层面的社会经济因素[9]。高校中数据中心的数据具有来源丰富、数据形式多样的特征，可开展如下研究：①在关于显性数据的研究中，消费数据、学业数据等一系列具有明显特征的数据可以更好地被观察，或利用统计学方法，将两种或多种看似不相关的变量联结起来，发现其蕴含的深层相关性。②在关于隐性数据的研究中，如学生的社交行为往往不能被直接观察，也不能通过简单的推理直接得到，这就需要利用如机器学习等数据分析手段实现。在国外的研究中常常引入隐性数据或隐性知识的概念，以解决企业运营和组织创新等问题[10]。有学者以中国大学生为研究对象，对其社交数据挖掘进行情感分析，深入观察学生的情感演化过程[11]。③在多源数据的研究中，显性数据和隐性数据可以综合起来，舒江波等就从学生学籍信息、学习表现、校园生活 3 个维度进行综合分析，构建学生大数据行为分析模型[12]。

2.3 不同应用场景的相关研究

当前的高校数据挖掘研究，受现实条件限制，开展特定场景中特定用户研究是可行的。国外的研究中也有利用混合数据对学生毕业情况进行专门统计，提出一种确定大学毕业状态驱动因素的公正方法。在国内研究中，由于教育体制不同，应用场景也有所不同：①在消费行为识别研究中，通过分析校园一卡通的消费数据，研究学生的消费行为，可以识别不同消费行为的群体[13]。②在贫困资助评估研究中，有学者在现

chinaXiv:202303.10408v1

有消费数据的基础上，对学生的发展状况进行调查，建立了一种贫困生资助评估模型，为识别和帮扶高校贫困生提供了新方法[14]；也有学者关注消费数据和学生个体的内在关联，提出一种用于消费强度指标，在学生家庭经济状况评估上进行了更为精准的预测[15]。③在心理健康评价研究中，由于心理相关的学生数据属于隐性数据，不能通过单一数据直接观察学生的心理状况。因此，学者大多采用多数据融合的方式，利用深度学习算法，构建大学生心理健康评估模型，实现自动准确评估大学生心理健康状态[16]。④在学生学业帮扶研究中，一方面，通过采集学习、生活过程中产生的校园行为数据，利用大数据的手段，可以构建面向学生的大数据分析模型，预测学生在校期间的学业表现[17]；另一方面，数据驱动的精准化学习评价可以发现教育教学中存在的问题，辅助课堂教学开展[18]。

这些研究既有基于显性数据、隐性数据的挖掘，也有基于多源数据融合的挖掘，但数据挖掘的深度仍然不够，缺乏对多源数据的深层挖掘。覆盖了多种应用场景，但仍然缺乏面向多场景的研究方法，虽然用户画像的提出可以解决场景单一的问题，但目前对学生画像的刻画上仍停留于框架的搭建，实践层面的学生画像研究成果较少，仍有一定的研究空白。因此，本文以大学生行为研究为出发点，获取真实的大学生的校园数据，通过将多源数据进行融合，构建多源、多维、多场景的综合评价体系。以消费、学业、社交 3 个维度构建动态和静态的个体画像。以消费维度研究为主，建立学生的消费活跃度和稳定性画像。其中，融合的优势在于数据、场景、深度的多元融合，最终刻画真实的、智能的、多层次的学生画像。基于高校学生画像，可以实现精准的群体圈选和个体识别，为高校贫困助学、学业帮扶和心理干预等工作提供参考，从而为高校管理提供理性决策依据。

3 高校学生画像的特征分析

3.1 学生特征分析

大学生既是具有独立意义的个体，也是具有社会意识的群体。在诸如高校此类小型社会系统中[19]，学生在校园学习、生活的同时，会建立以自我为核心的社交网络，在范围上，既有以寝室、专业、班级为单位的自然社交网络，也有跨年级、跨学院、跨角色的主观社交网络。在学生进行社交活动的过程中，根据不同粒度的用户行为特征可以划分出很多不同种类的用户角色，学生既可以是“有影响力者”“专家”或“讨论者”，也可以是“支持者”“中立者”或“反对者”。但是，学生无论扮演何种角色，都会在其社交网络中发挥影响。由此可见，学生的校内行为数据具备个体和群体的双重数据特征，反映真实的个人特征和社交关系，在研究中，既要重视学生的个体性，又不能忽视学生的群体特点。

3.2 学生画像的属性特征

从宏观角度来看，学生画像的属性特征兼具静态性和动态性。从行为层面来看，可以把学生的在校行为划分为学习行为、消费行为和社交行为 3 类。

（1）学业行为指标。学业行为指标主要包括学业成绩优秀度、学业努力程度等。在教育领域，对于学生的学习评价方式有很多，目前各高校普遍根据学生的培养方案课程，以学分作为权重计算学生学分绩点，部分学者提出以专业排名作为评价学业优秀度的评价标准[20]。在评价学习行为的过程中，要根据学生学制、学年、专业的不同分类评价，并结合如奖学金、竞赛等学科竞赛信息和图书馆出入信息，研究学生的学习努力程度，构建客观、合理、简洁的学业评价指标。

（2）消费行为指标。消费行为指标主要包括消费稳定性、消费活跃度、消费水平等。高校为在校师生提供了基础的生活需求保障，因此，通过研究校园内学生的消费行为，包括学生的消费时间、金额、地点信息，进一步可以形成消费时间稳定性和消费地点偏好等指标，并在一定程度反映了学生参与校内活动，融入校园生活的实际情况。

（3）社交行为指标。社交行为指标主要包括社交活跃度和社交距离度，受研究规模影响，高校属于小型的社会系统，在高校范围内开展社交距离度的研究

chinaXiv:202303.10408v1

意义不大。因此，可以将社交活跃度近似看作社交行为指标。通过追踪学生的消费数据，建立消费“时间－地点”共现网络，发现异常离群值，甄别学生群体中的“离群者”，实现社交行为指标的确定。

3.3 学生画像的数据特征

基于学生群体特征及画像的属性特征所构建的学生画像的数据特征具备客观性、全面性、融合性和动态性[21]。其中，客观性是指学生画像基于一系列真实数据构建，符合个体和群体层面的实际状况，反映真实科学的属性特征，数据来源客观、处理手段客观、研究目的客观、呈现方式客观；全面性是指学生画像构建涉及学生行为的全方面，也反映了学生特征的全方面，具体体现在研究角度和业务场景的全覆盖；融合性是指各职能部门的异构数据相互融合，实现数据融合时要求完整融合、按属性融合、按业务场景融合；动态性是指用户画像具有动态变化的特征，个体在不同时期所表现的特征不同，导致刻画的用户画像也有所差异，因此学生画像也是一个实时变化的动态模型。

3.4 总体框架设计

高校学生用户画像的数据来源为教务管理部门、学生管理部门、一卡通中心、图书馆等职能部门，整个研究大致分为 3 个层级：数据层、挖掘层和表征层，如图 1 所示。①数据层。包括基本信息数据、教务成绩数据、奖助学金数据、图书馆门禁记录和校园消费数据。获取多源异构数据后，进行清洗、集成、转换和规约，完成数据融合。②挖掘层。主要是对预处理后的数据进行指标分析、聚类分析、相关性分析和共现分析，然后建立关于学生的消费行为指标、学业行为指标和社交行为指标的标签集，建立个体画像和群体画像。③表征层。利用学生个体画像实现学业预警、心理预警和贫困帮扶，利用学生群体画像实现重点群体识别、群体行为预测和校园资源规划等方面的应用表征。

4 研究过程

4.1 数据采集

本实验选取黑龙江省某高校 2018 级、2019 级在校生 2019—2020 年的学生日常行为记录数据作为数据集，利用 MySQL 导出数据 40 余万条。包括基本信息数据、教务成绩数据、奖助学金数据、图书馆门禁记录和校园消费数据 5 张数据表，基本情况如表 1 所示。

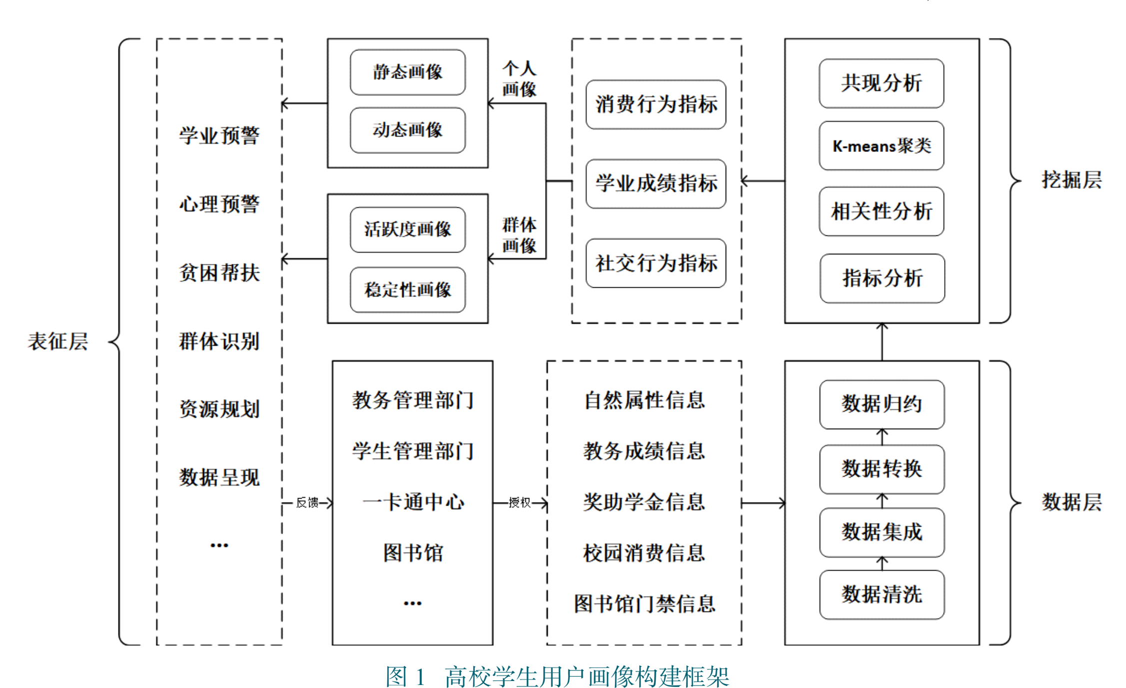


图 1　高校学生用户画像构建框架

Fig.1 Construction framework of college student user profiles

ChinaXiv合作期刊

表 1　学生基本数据

Table 1　Basic student data

授权部门	表名	列名
学生工作部	基本信息数据	学号、年级、专业、学制、学籍状态
	奖助学金数据	学号、年级、专业、奖学金名称、奖学金级别、获奖时间、助学金名称、助学金级别、获助时间
教务处	教务成绩数据	学号、年级、专业、课程名称、课程类别、学分、成绩
图书馆	图书馆门禁数据	学号、卡号、刷卡时间、闸机号
一卡通中心	校园消费数据（一卡通）	学号、消费时间、消费金额、消费后余额、消费地点、消费商户名称
	校园消费数据（在线支付）	学号、消费时间、消费金额、消费后余额、消费地点、消费商户名称、消费类别（微信、支付宝）

4.2　数据预处理

各个部门授权的数据多为结构化数据，将授权后的数据导入到 SPSS 进行处理，清洗部分格式不规范或错误的数据后，将数据表以“学号”字段作为特征匹配项进行数据融合，保留以“学号”为字段的研究对象 593 个，时间范围为 2019 年 3 月至 2020 年 12 月，共 4 个学期。其中，受新冠肺炎疫情影响，2020 年上半年未正常开展线下教学工作，因此 2020 年上半年的消费记录不计入研究范围。

4.3　数据分析与特征提取

4.3.1　学生消费行为特征

根据“消费地点”字段，可以将消费数据按“日常生活”“健身洗浴”“基本饮食”和“健康医疗”分类。根据“消费地点”字段，结合校园内商户的分布情况，将消费数据的地点按“A 区”“B 区”和“C 区”分类。在“基本饮食”分类下，结合食堂的实际开放时间和就餐高峰人数统计，划分“6:00—9:30”为早餐时间、“10:30—14:00”为午餐时间、“16:30—20:00”为晚餐时间，并将同一时间段内的多笔消费合并为一笔。

经过征集学生的消费习惯，并结合学校实际情况。学生在校园内的饮食与购物行为习惯基本一致，且“基本饮食”支出比重较大，可以将就餐行为近似视作学生的消费行为。因此，本研究中学生的“消费行为”数据按“就餐行为”数据计算。

就餐时间稳定系数是对学生就餐时间稳定性的描述，记为 λ，如公式（1）所示：

$$\lambda = \frac{\sum_{i=1}^{n} MT_{sd_i} \times N_{m_i}}{\sum_{i=1}^{n} N_{m_i}} \tag{1}$$

其中，MT_{sd_i} 表示第 i 个餐别就餐时间的标准差，其计算方法如公式（2）所示；N_{m_i} 表示第 i 个餐别就餐总次数；n 表示餐别种类，本文取值为 3。

$$MT_{sd_i} = \sqrt{\frac{1}{N}\sum_{j=1}^{n}(T_j + \overline{T})^2} \tag{2}$$

其中，N 表示某个餐别的就餐总次数；T_j 表示某个餐别的第 j 次就餐时间；$\overline{T}$ 表示某个餐别的平均就餐时间。

4.3.2　学生学业行为特征

学生的学业行为特征主要由学业优秀度评价，同一年级、同一专业的学生成绩排名越高，其学业优秀度也就越高。学业优秀度是对学生学业成绩的优秀程度的描述，记为 σ，如公式（3）所示。

$$\sigma = \frac{G - G_{min}}{G_{max} - G_{min}} \tag{3}$$

G 表示学生的学分绩点，如公式（4）所示；G_{max} 表示某学生所在专业最高成绩，G_{min} 表示某学生所在专业最低成绩。

$$G = \frac{\sum_{i=1}^{n} G_i \times F_i}{10 \times \sum_{i=1}^{n} F_i} \tag{4}$$

其中，G_i 表示某学生在第 i 门课程中的期末成绩；F_i 表示某学生第 i 门课程的学分值；n 表示某学生年度选修的课程总数。

此外，学生的学业行为特征包括学业努力程度评价，而学业努力程度评价可以通过获取在学习行为上

chinaXiv:202303.10408v1

chinaXiv:202303.10408v1

付出的时间计算得出，主要体现为一个学期内学生进出图书馆的有效次数，但学生进出图书馆次数并不与学业行为直接相关，只能作为学业行为特征的辅助评价指标。

4.3.3 学生社交行为特征

好友关系是学生社交行为的重要体现，是学生社交网络研究的主要内容。学生往往会和同寝室与同班级的好友一起出行，如果两个人多次在同一时间段、同一地点存在消费行为，且共现的概率值大于一定的阈值时，则认为两个人存在好友关系。在已有的关联规则基础上，借鉴已有学者的共现网络算法，假设学生 X 在某一时刻进行食堂刷卡消费行为，在一定的时间间隔内，学生 Y 也在同一消费地点出现刷卡消费行为，则认为学生 X 与 Y 存在共现行为，当关联规则 XY 满足最小支持度和最小置信度阈值时，认为学生 X 和学生 Y 之间存在关联，即认定两人为好友关系。

在社交共现分析中，学生 X 和学生 Y 的好友关系反映到数据层面，可以理解为学生 X 和学生 Y 共现的次数足够大，且共现的消费记录占自身所有消费记录较大比重。因此，设置最小置信度为 $\beta=0.5$，最小支持度 α 如公式（5）所示。

$$\alpha=\frac{R}{10N} \tag{5}$$

N 表示所有刷卡消费的学生数，R 表示所有学生的刷卡消费记录总数。

为计算学生 X 和学生 Y 好友关系的可能性，引入置信度 $C_{X\to Y}$ 如公式（6）所示。

$$C_{X\to Y}=\frac{S_{X\to Y}}{S_X} \tag{6}$$

其中，$S_{X\to Y}$ 为学生 X 和学生 Y 的共现次数，S_X 为学生 X 刷卡消费的总次数。

在对学生 X 和学生 Y 的好友关系判定过程中，首先，计算学生 X 和学生 Y 的共现次数 $S_{X\to Y}$，若 $S_{X\to Y}\geqslant\alpha$，则说明两人的共现次数足够高；下一步，则计算学生 X 和学生 Y 的好友可能性置信度 $C_{X\to Y}$，若 $C_{X\to Y}\geqslant\beta$，则说明两人存在好友关系。

5 高校学生画像构建与呈现

5.1 学生个体画像

在学生个体画像的构建中，通过对消费、社交和学业数据的指标进行分类，获取画像标签，可以实现学生整体状况的观测。利用 MySQL 数据库完成数据清洗，SPSS 对数据进行处理与分析，获取学生有关学业行为、消费行为和社交行为的 3 类指标。本研究选取学生 A 作为案例，如表 2 所示。其标签信息加载到学生个体画像模型，如图 2 所示。其中，“值”内的文本部分为画像的分类属性，根据学生的排名位次分类得到。

（1）在学业画像中，整体上看，该生学业成绩优秀，在学业成绩位于同专业前列，数据表示前往图书馆的次数较多，学业努力程度和学业优秀度都很高，且没有任何违纪处分，可以推测该生具有较强的自主学习能力和自我约束力，同时验证了学业努力程度与学业优秀度存在一定的正相关关系。

（2）在消费画像中，该生表现出较强的消费稳定性和消费活跃性，总消费次数较高，常常使用在线支付的方式，初步推测平时校内生活较为丰富。此外，在消费地点的选择上，学生的消费记录在 A 区较多，推测该生的校内活动受一定时空因素的限制，或受个人主观因素影响，在校内活动时轨迹较为集中。另一方面，该生的就餐时间集中在中午较多，在早上的就餐支出较少，消费不稳定，就餐缺乏规律，推测缺少健康的饮食习惯。

（3）在社交画像中，该生的社交评价为优秀社交，初步认定该生拥有良好的社交关系，具备一定的社交能力和团体意向，进一步推测此学生现阶段处于心理健康积极的状态，在生活中遇到困难时会更易得到好友的帮助。

综上，该生呈现出学业优秀、消费活跃、社交良好的应届毕业生形象，结合学业、消费和社交 3 个维度的综合评估，该生属于高活跃的校园生活者，为人努力上进，心理健康向上，虽然在消费（就餐）规律

表 2 学生画像标签信息

Table 2 Student profile label information

类别	标签	解释	值	属性
自然属性	性别	自学生入学起，不随时间更改的自然属性指标	男、女	静态
	年级		-	静态
	专业背景		-	静态
学业行为标签	学业优秀	关于学生学习行为的结果性评价指标，即该生在同一年级、同一专业内的学习成绩排名	0%~10%：优秀 10%~30%：良好 30%~60%：合格 60%~100%：不合格	动态
	学业努力	关于学生学习行为的过程性评价指标	0%~10%：非常努力 10%~30%：很努力 30%~60%：努力 60%~100%：尚需努力	动态
	违纪行为	关于学生学习真伪性的辅助评价指标	有违纪、无违纪	动态
消费行为标签	消费水平	关于学生消费水平的评价指标，即该生的总消费金额在样本群体中的相对水平	0%~33%：高水平 33%~66%：中水平 66%~100%：低水平	动态
	消费次数	关于学生消费次数的评价指标，即该生的总消费次数在样本群体中的相对水平	0%~33%：高次数 33%~66%：中次数 66%~100%：低次数	动态
	消费规律	关于学生消费规律的评价指标，即该生的不同时段消费时间的波动程度	时间稳定系数大于样本群体均值：稳定 时间稳定系数小于样本群体均值：不稳定	动态
	时间偏好	关于学生消费行为的时间偏好，即该生有效消费次数最高的时段	早上、中午、下午	动态
	地点偏好	关于学生消费行为的地点偏好，即该生有效消费次数最高的地理位置区域	A 区、B 区、C 区	动态
	支付偏好	关于学生消费行为的支付偏好，即该生有效消费次数最高的支付方式	一卡通支付、在线支付（含支付宝和微信）	动态
社交行为标签	社交评价	关于学生社交行为的评价指标，即该生可疑好友数量在样本群体中的相对水平	可疑好友数大于样本群体均值：优秀社交 可疑好友数小于样本群体均值：加强社交	动态

自然属性	性别	年级	专业背景
	女	2017 级	图书馆学
学业行为标签	学业优秀	学业努力	违纪行为
	优秀	努力	无违纪
消费行为标签	消费水平	消费次数	消费规律
	高水平	高次数	不稳定
	时间偏好	地点偏好	支付偏好
	中午	*A* 区	在线支付
社交行为标签	社交评价		
	优秀社交		

图 2 学生 *A* 的学生个体画像标签信息

Fig. 2 Personal profile labels of student *A*

上呈现不稳定的状态，但是整体还是自律的学生。由于学生处于大四毕业期间却仍有高度的学业努力度，可以初步预测学生有求学备考或求职复习的准备，学校针对此类学生可以提供针对的信息推送服务或安排对应的辅导课程。

5.2 学生群体消费画像

5.2.1 基于消费活跃度的群体画像

本文主要采用 K-means 聚类方法对学生行为特征进行聚类[22]。利用 Python 中 sklearn 工具实现 K-means

chinaXiv:202303.10408v1

chinaXiv:202303.10408v1

聚类算法，对学生的“就餐天数”“就餐金额”进行聚类，以探究使学生用餐行为的共性群体，实验过程中，随着聚类数 k 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，因此，利用手肘法可以确定 k 值的继续增大而趋于平缓的拐点。如图 3 所示，发现当 k=2 时的聚类效果较好，聚类中心的各项特征数据值如表 3 所示。

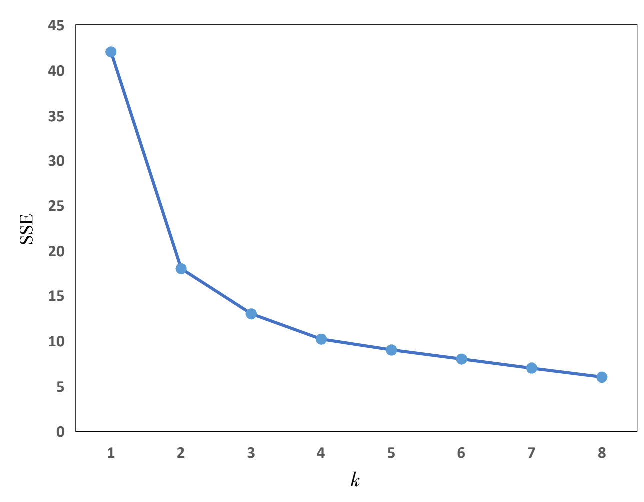


图 3　“就餐天数”“就餐金额”聚类不同 k 取值

Fig.3 "Dining days" and "dining expenditure" clusters with different k values

表 3　就餐行为聚类中心

Table 3 Dining behavior cluster centers

聚类特征	类别 1	类别 2
就餐天数/天	91.28	64.78
就餐金额/元	1 757.39	933.86

在根据就餐行为聚类中心结果中，通过对学生的“就餐天数”“就餐金额”进行聚类，可以有效衡量学生的消费活跃度和校园活跃度。其中，类别 1 的学生有 216 人，占比为 36.42%；类别 2 的学生有 377 人，占比为 63.58%。

类别 1 的学生消费天数较多，消费金额也明显高于其他聚类中心，处于此类别的学生属于消费活跃度高的群体，他们在学校消费的天数和金额都很高。此外，不仅在消费活跃度上，在校园生活中也表现出极高的活跃度，属于校园生活的重要参与者。往往这类学生都比较关注学校相关政策和服务设施的变化，在学校开展校园意见征集时，此类学生的意见将具备一定的参考性。此外，在此类消费活跃度高的学生中，会存在消费天数高于聚类中心，且消费金额低于聚类中心的情况，此类学生的日常饮食都会在食堂进行，而且单次消费水平较低，可以考虑是否存在贫困情况，学校也应重点关注此类学生，为其日常生活提供必要及时的保障。

类别 2 的学生消费天数和消费金额都处于中等水平，也是占全体学生较大比例的一部分群体。这些学生消费活跃度适中，无法通过就餐天数和就餐金额判断学生的贫困情况，可以结合学生的助学金申请情况，将消费活跃度适中，但就餐天数远高于聚类中心的学生判定是否为贫困生，为学校的助学工作提供参考。

5.2.2　基于消费稳定性的群体画像

对学生的“早餐就餐率”“午餐就餐率”和“晚餐就餐率”进行聚类，实验过程中，不断调节 k 值分别进行实验对比，发现当 k=3 时的聚类效果较好，如图 4 所示，聚类中心的各项特征数据值如表 4 所示。

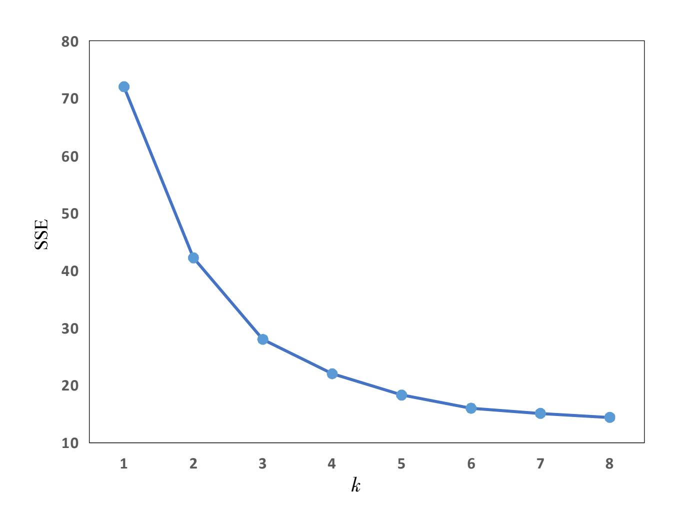


图 4　“早餐就餐率”“午餐就餐率”和“晚餐就餐率”聚类不同 k 取值

Fig.4 "Breakfast rate", "lunch rate" and "dinner rate" clusters with different k values

在根据就餐行为聚类中心结果中，通过对学生的“早餐就餐率”“午餐就餐率”和“晚餐就餐率”进行聚类，可以有效衡量学生的消费稳定性和自律性。其中，类别 1 的学生有 65 人，占比为 10.96%；类别 2

表 4 就餐规律聚类中心

Table 4 Clustering centers of dining patterns

聚类特征	类别 1	类别 2	类别 3
早餐就餐率	0.499	0.171	0.084
午餐就餐率	0.654	0.403	0.157
晚餐就餐率	0.561	0.242	0.087

的学生有 209 人，占比为 35.25%；类别 3 的学生有 319 人，占比为 53.79%。

类别 1 的学生三餐就餐率都很高，和其他聚类中心相比，此类别的学生一般都有着健康的饮食习惯，在生活习惯上反映出较强的自律性。类别 2 的学生午餐就餐率较高，但早餐和晚餐就餐率较低，此类别的学生通常就餐不规律，早餐就餐率较低的学生通常早起率也很低，缺乏生活习惯上的自我约束；晚餐就餐率较低的学生考虑存在节食的情况，应当鼓励此类学生养成健康的饮食习惯，形成科学规律的生活作息。类别 3 的学生三餐就餐率都很低，此类学生同样存在校内活动少的情况，存在校外就餐和订外卖的情况，无法通过校园消费数据准确推测其生活习惯。

5.3 高校学生画像的应用表征

基于多源数据融合的高校学生画像构建，结合学生三维行为特征，可以分别构建学生个体画像和学生群体画像。针对面向的业务场景不同，学生画像也有着不同方面的应用表征。

（1）学生异常识别与预警。通过对学生个体画像的观测，可以对学生的消费、学业和社交 3 个方面进行初步评估，发现在学生画像中表现出的优势值，为评奖评优工作提供参考，为助学助困工作提供证明。对学业努力且学业优秀，但违纪次数异常值的发现，方便及时安排重点关注及谈心谈话。此外，通过对学生画像动态观测，对比变化及时预警，有利于学生工作部门和辅导员发现存在的学业和心理问题，及时帮助学生应对在思想取向、价值引领、学习生活、择业交友等方面的具体问题。

（2）学生群体关注与引导。基于聚类算法的学生群体画像构建，聚焦于学生的消费行为，发现学生的典型特征区分，在消费稳定性和活跃度上表现出明显的群体属性。在消费活跃画像的结果分析中，学生被分类成典型的高活跃和低活跃两个群体，给予低活跃群体更多关注。同理，在消费稳定性的结果分析中，学生被分类成高稳定、中稳定和低稳定 3 个群体。在实际的学生管理工作中，学生工作部门和辅导员应当更多关注低活跃和低稳定群体，发现学生存在的潜在不良消费习惯和饮食习惯，尤其是在疫情防控管理期间，对校内消费画像进行观察，更好的预判校内与社会面的接触风险，对相关学生进行及时有效地引导和规劝。

（3）校园资源规划与调节。结合学生个体画像和群体画像的结果，学生的早晚餐习惯状况欠佳。为养成良好的消费习惯和用餐习惯，可以利用学生画像对校内资源规划进行预判和规划，如为学生消费较多的校区开设更多的就餐座位，延长就餐时间，减轻高峰就餐压力。在消费较少的校区开设特色餐厅，引导学生分布就餐，利用分流缓解就餐压力。另外，为提高学生早晚就餐率及就餐稳定性，学校可以推出更多种类餐品，配合开展健康饮食习惯普及宣传活动，帮助学生养成良好的就餐习惯，实现资源的科学、合理、人性规划，为调节学校资源分配和决策提供具体参考。

6 结　语

本文以高校数据化管理为研究背景，对高校数据挖掘研究进行以下创新。首先，本文尝试利用一种新的数据融合视角，通过将显性数据与隐性数据融合，并生成有关消费行为、学业行为和社交行为三维指标。其次，为了解决以往研究中应用场景单一问题，现利用用户画像的手段，实现多场景的融合。最后，本研究基于学生的真实数据，在以往学生画像的研究基础上，利用 SPSS 和 K-means 聚类算法等方法，圈选不同特征的学生群体，同时利用学生共现网络，研究学生的社交关系，对某高校学生数据进行分析，进一步进行了实证研究，刻画大学生的“消费 - 学业 - 社交”画像。在多源数据融合视角下构建学生画像，可以有效

chinaXiv:202303.10408v1

ChinaXiv合作期刊

chinaXiv:202303.10408v1

为高校教务、学工等部门决策提供依据，尤其是后疫情时代对大学生画像可以及时发现潜在的风险隐患。研究分析发现：①在学生个体画像中，通过对学生画像标签信息的解读，可以对学生消费、学业和社交 3 个方面的情况进行了解，实现学生个体的动态监测；②在学生群体画像中，通过聚类分析的方法，可以圈选不同特征的学生群体，尤其是在消费行为方面，深度分析学生的活跃度和稳定度特征，既可以为宏观层面的学生观测提供依据，又为探寻学生不同行为要素间的相关性提供了新的思路；③在应用表征层面，融合多场景的学生画像可以同时实现高校异常识别与预警、群体关注与引导和资源规划与调节，大大拓宽了研究的应用场景，提升高校教育教学管理能效。

在大数据时代下，信息化的高校管理已成为当代的研究重点，为了更好地实现高效、智能、多元化管理，学生画像提供了一种新的研究思路。但受数据、算法的局限性，学生画像的准确性和易用性还有待提高，既有现实条件的约束，也有研究手段的不足，在未来的研究中，应通过更广地调研研来完善大学生画像构建体系，并不断尝试改进更为合适的画像技术，将高校学生画像应用到更多业务场景中。

参考文献：

[1] 刘邦奇, 袁婷婷, 纪玉超, 等. 智能技术赋能教育评价: 内涵、总体框架与实践路径[J]. 中国电化教育, 2021(8): 16–24.
LIU B Q, YUAN T T, JI Y C, et al. Intelligent technology enabling education evaluation: Connotation, overall framework and practice path[J]. China educational technology, 2021(8): 16–24.

[2] 梁樱花. 大数据对我国高校教育管理的影响及其应对措施——评《基于大数据的高校教育管理研究》[J]. 中国科技论文, 2021, 16(11): 1287.
LIANG Y H. The influence of big data on the education management of colleges and universities in my country and its countermeasures – Comment on "research on college education management based on big data"[J]. China sciencepaper, 2021, 16(11): 1287.

[3] 朱东妹. 多源数据融合视角下的阅读推广用户画像构建研究[J]. 图书馆理论与实践, 2021(6): 99–105.
ZHU D M. Research on the construction of user profile for reading promotion from the perspective of multi –source data fusion [J]. Library theory and practice, 2021(6): 99–105.

[4] ZAUGG H, RACKHAM S. Identification and development of patron personas for an academic library[J]. Performance measurement and metrics, 2016, 17(2): 124–133.

[5] 张治, 刘小龙, 徐冰冰, 等. 基于数字画像的综合素质评价: 框架、指标、模型与应用[J]. 中国电化教育, 2021(8): 25–33, 41.
ZHANG Z, LIU X L, XU B B, et al. Comprehensive quality assessment based on digital portrait: Framework, indicators, model and applications[J]. China educational technology, 2021(8): 25–33, 41.

[6] 余明华, 张治, 祝智庭. 基于可视化学习分析的研究性学习学生画像构建研究[J]. 中国电化教育, 2020(12): 36–43.
YU M H, ZHANG Z, ZHU Z T. Research on the construction of student portrait in research – Based learning based on visual learning analytics[J]. China educational technology, 2020(12): 36–43.

[7] 邓嘉明. 智慧校园学生数据画像生成方式研究[J]. 现代电子技术, 2019, 42(21): 58–62.
DENG J M. Research on creation ways of data image of students in intelligent campus[J]. Modern electronics technique, 2019, 42 (21): 58–62.

[8] 邹丽伟, 刘晋禹. 智慧育人理念下的大学生信息精准服务研究[J]. 情报科学, 2021, 39(8): 120–125.
ZOU L W, LIU J Y. Accurate information service for college students under the concept of intelligent education[J]. Information science, 2021, 39(8): 120–125.

[9] GILLEN–O'NEEL C, ROEBUCK E C, OSTROVE J M. Class and the classroom: The role of individual– and school–level socioeconomic factors in predicting college students' academic behaviors[J]. Emerging adulthood, 2021, 9(1): 53–65.

[10] M UTHUVELOO R, SHANMUGAM N, TEOH A P. The impact of tacit knowledge management on organizational performance: Evidence from Malaysia[J]. Asia pacific management review, 2017, 22 (4): 192–201.

[11] ZHAO H, ZUO Y, XU C, et al. What are students thinking and feeling? Understanding them from social data mining[J]. International journal of computer applications in technology, 2021, 65 (2): 110–

117.

[12] 舒江波, 葛雄, 彭利园, 等. 基于学生个人大数据的行为特征分析[J]. 华中师范大学学报(自然科学版), 2020, 54(6): 927-934.

SHU J B, GE X, PENG L Y, et al. Analysis of behavioral characteristics based on student's personal big data [J]. Journal of central China normal university(natural sciences), 2020, 54(6): 927-934.

[13] 龚黎旰, 顾坤, 明心铭, 等. 基于校园一卡通大数据的高校学生消费行为分析[J]. 深圳大学学报(理工版), 2020, 37(s1): 150-154.

GONG L G, GU K, MING X M, et al. Analysis of college students' consumption behavior based on campus card data[J]. Journal of Shenzhen university(science and engineering), 2020, 37(s1): 150-154.

[14] 张存禄, 马莉萍, 陈晓宇. 贫困生资助对大学生消费行为的影响: 基于校园卡消费大数据和问卷调查数据的研究[J]. 教育与经济, 2021, 37(3): 80-87, 96.

ZHANG C L, MA L P, CHEN X Y. A study of the mobility of rural young and middle-aged teachers: Field, habitus, and capital theory[J]. Education & economy, 2021, 37(3): 80-87, 96.

[15] 宋德昌. 基于校园卡的学生经济状况评价方法研究[J]. 中山大学学报(自然科学版), 2009, 48(s1): 9-11.

SONG D C. Research on the evaluation method of student economy status based on campus card[J]. Acta scientiarum naturalium universitatis sunyatseni, 2009, 48(s1): 9-11.

[16] 周炫余, 刘林, 陈圆圆, 等. 基于多模态数据融合的大学生心理健康自动评估模型设计与应用研究[J]. 电化教育研究, 2021, 42(8): 72-78.

ZHOU X Y, LIU L, CHEN Y Y, et al. Research on design and application of an automatic assessment model for college students' mental health based on multimodal data fusion [J]. E-education research, 2021, 42(8): 72-78.

[17] 张华, 刘颖. 运用多任务排序学习算法预测学业成绩[J]. 扬州大学学报(自然科学版), 2020, 23(5): 63-67.

ZHANG H, LIU Y. Predicting academic performance using multi-task learning RankNet[J]. Journal of Yangzhou university(natural science edition), 2020, 23(5): 63-67.

[18] 黄涛, 赵媛, 耿晶, 等. 数据驱动的精准化学习评价机制与方法[J]. 现代远程教育研究, 2021, 33(1): 3-12.

HUANG T, ZHAO Y, GENG J, et al. Evaluation mechanism and method for data-driven precision learning[J]. Modern distance education research, 2021, 33(1): 3-12.

[19] NIE M, YANG L, SUN J, et al. Advanced forecasting of career choices for college students based on campus big data[J]. Frontiers of computer science, 2018, 12(3): 494-503.

[20] QU S, LI K, ZHANG S, et al. Predicting achievement of students in smart campus[J]. IEEE access, 2018, 6: 60264-60273.

[21] ZHANG W, JIANG L. Algorithm analysis for big data in education based on depth learning [J]. Wireless personal communications, 2018, 102(4): 3111-3119.

[22] 王卫芳. 基于校园大数据的学业表现预测及行为分析[D]. 重庆: 重庆大学, 2019.

WANG W F. Academic performance prediction and behavior analysis based on campus big data[D]. Chongqing: Chongqing university, 2019.

chinaXiv:202303.10408v1

chinaXiv:202303.10408v1

Construction of College Students' "Consumption-Academic-Social" Profiles from the Perspective of Multi-source Data Fusion

HUANG Taihua[1], ZHANG Tao[1*], WANG Lei[2]
(1. School of Information Management, Heilongjiang University, Harbin 150080; 2. School of Data Science and Technology, Heilongjiang University, Harbin 150080)

Abstract: [Purpose/Significance] Mining college student data and constructing studnet profiles is conducive to in-depth understanding of students' needs, improving management level, and promoting intelligent service. [Method/Process] Based on the multi-source data mainly generated by the management and service process of colleges and universities, student profiles were developed by focusing on consumption, academic and social indicators, analyzing the characteristics of students, using the Scikit-Learn tool of Python, and applying the K-means clustering algorithms. Empirical research was carried out and representativeness of student portraits from individual and group perspectives was studied. [Results/Conclusions] First, this paper attempts to utilize a new data fusion perspective, by fusing explicit data with implicit data, and generating three-dimensional indicators of consumption behavior, academic behavior, and social behavior. Secondly, in order to solve the problem of single application scenario in previous research, the method of user profile construction is used to realize the fusion of multiple scenarios. Finally, based on the real student data, this study uses K-means clustering algorithm to select groups of students with different characteristics on the basis of previous research. The data of college students is analyzed, and further empirical research is carried out to describe the "consumption-academic-social" profiles of college students. Constructing student profiles from the perspective of multi-source data fusion can effectively provide a basis for decision-making by different units in colleges and universities, such as academic affairs,. Especially in the post-epidemic era, the profiles of college students can detect potential risks in time. The study found that at the individual level, by interpreting the label information of students' portraits, it is possible to understand the 3 aspects of students' consumption, academics and social interaction, and realize dynamic monitoring of individual students. At the group level, through cluster analysis, students with different characteristics can be selected, especially in terms of consumption behavior, and the characteristics of students' activity and stability can be deeply analyzed, which can not only provide a basis for the macro-level observation of students, but also provide new ideas for exploring the correlation between different behavioral elements of students. At the application level, the integration of multi-scenario student profiles can simultaneously realize abnormal identification and early warning, group attention and guidance, and resource planning and adjustment, which greatly broadens the application scenarios of research and improves the energy efficiency of education and teaching management in colleges and universities. However, due to the limitations of data and algorithms, the accuracy and ease of use of student portraits still need to be improved. There are both constraints from practical conditions and insufficient research methods. In future research, more extensive research should be used to improve college student profile construction system, and constantly develop more suitable techniques.

Keywords: student profile; consumption analysis; social analysis; academic analysis; K-means clustering; information behavior